Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 32(10): 5783–5803.
DOI: 10.3934/era.2024267
Received: 24 July 2024
Revised: 09 September 2024
Accepted: 24 September 2024
Published: 24 October 2024

Research article

MFCEN: A lightweight multi-scale feature cooperative enhancement network for single-image super-resolution

Jiange Liu[1], Yu Chen[2], Xin Dai[1], Li Cao[1] and Qingwu Li[2,*]

[1] State Grid Huaian Power Supply Company, Jiangsu 223001, China

[2] College of Information Science and Engineering, Hohai University, Jiangsu 213200, China

* **Correspondence:** Email: li_ qingwu@163.com.

Abstract: In recent years, significant progress has been made in single-image super-resolution with the advancements of deep convolutional neural networks (CNNs) and transformer-based architectures. These two techniques have led the way in the field of super-resolution technology research. However, performance improvements often come at the cost of a substantial increase in the number of parameters, thereby limiting the practical applications of super-resolution methods. Existing lightweight super-resolution methods, which primarily focus on single-scale feature extraction, lead to the issue of missing multi-scale features. This results in incomplete feature acquisition and poor reconstruction of the image. In response to these challenges, this paper proposed a lightweight multi-scale feature cooperative enhancement network (MFCEN). The network consists of three parts: shallow feature extraction, deep feature extraction, and image reconstruction. In the deep feature extraction part, a novel integrated multi-level feature module was introduced. Compared to existing CNN and transformer hybrid super-resolution networks, MFCEN significantly reduced the number of parameters while maintaining performance. This improvement was particularly evident at a scale factor of 3. The network introduced a novel comprehensive integrated multi-level feature module, leveraging the strong local perceptual capabilities of CNNs and the superior global information processing of transformers. It was designed with depthwise separable convolutions for extracting local information and a block-scale and global feature extraction module based on vision transformers (ViTs). While extracting the three scales of features, a satisfiability attention mechanism with a feed-forward network that can control the information was used to keep the network lightweight. Experiments demonstrated that the proposed model surpasses the reconstruction performance of the 498K-parameter SPAN model with a mere 488K parameters. Extensive experiments on commonly used image super-resolution datasets further validated the effectiveness of the network.

Keywords: single-image super-resolution; lightweight; multi-scale; attention mechanism

1. Introduction

Single-image super-resolution (SISR) represents a crucial branch in the field of image restoration, aiming to recover a high-resolution image from a low-resolution counterpart. Widely applicable in defense [1], military [2], medical imaging [3, 4], and facial recognition domains [5], it not only enhances image perceptual quality [6–9] but also contributes to improving various machine vision tasks [10–12]. However, this inherent ill-posed nature of SISR [13], where multiple high-resolution images correspond to a single low-resolution image, has rendered it challenging over the past decades. Consequently, there is a growing focus on researching image super-resolution networks with robustness and high computational speed [14, 15].

Since Harris [16] first introduced the task of super-resolution reconstruction, mainstream methods can be categorized into three types: interpolation-based, reconstruction-based, and learning-based approaches [17]. Interpolation-based and reconstruction-based approaches are considered traditional super-resolution methods. In contrast, learning-based approaches have evolved significantly, leading to the emergence of many advanced networks. CNN-based and transformer-based methods have achieved superior reconstruction performance compared to traditional techniques. The application of convolutional neural networks (CNNs) to image super-resolution reconstruction, exemplified by the SRCNN network introduced by Dong [18], has not only ensured reconstruction image quality but also significantly improved reconstruction speed. This shift marked the superiority of CNN-based methods over traditional techniques such as bicubic interpolation. Subsequently, numerous CNN-based SISR models emerged, including FSRCNN [19], ESPCN [20], VDSR [21], RRSR [22], RDDAN [23], and NLSA [24], among others.

In recent years, the introduction of the Swin transformer [25], leveraging a sliding window mechanism for self-attention, has propelled vision transformers (ViTs) into the limelight in computer vision tasks [26, 27], achieving state-of-the-art performance in various advanced vision tasks [28]. However, ViTs exhibit certain challenges when applied to image super-resolution tasks. Due to the primary usage of transformers for sequence modeling, their self-attention mechanism is more adept at capturing global relations. Yet, in handling image details, the relatively smaller receptive field proves challenges in effectively capturing high-frequency details [29–31]. Additionally, the self-attention mechanism computes pairwise token affinities for all spatial positions, leading to high computational complexity and substantial memory usage for large-scale images [25, 32, 33]. The elevated parameter count may hinder real-time operation, especially on resource-constrained mobile devices. Thus, there is an urgent need to develop a method that can address multi-scale feature extraction and fusion while reducing the computational cost of reconstructing high-quality, high-resolution images.

In response to these challenges, this paper proposes a lightweight multi-scale feature cooperative enhancement network (MFCEN), with the following key contributions:

• Introduction of a local feature capture module utilizing depth convolution based on depthwise separable convolutions and a squeeze-and-excitation module (SE) to incorporate CNN-based local features into the transformer-based global features, enhancing the local perception capability of the network.

• Design of a block-scale feature extractor module, innovatively proposing a dynamic sparse self-attention mechanism for more lightweight and flexible feature extraction at the block scale.

• Introduction of an innovative global semantic capture module, incorporating a cross-channel self-

attention mechanism and a controlled information feed-forward network to effectively capture global information while reducing computational complexity and parameter count.

These modules construct a comprehensive multi-level feature module. This compensates for the limitations of existing transformer architectures in establishing cross-scale attention mechanisms, utilizing features from local, global, and block-scale enhanced structures. Adhering to the lightweight design philosophy, this method consistently balances performance with computational complexity. The proposed approach not only contributes to enhancing the performance of image super-resolution tasks but also provides an innovative solution for reducing computation costs. The remainder of this paper is organized as follows: Section 2 reviews related work, discussing existing methods in the field of the proposed multi-scale feature cooperative enhancement network (MFCEN), including its network architecture and key components. Section 4 presents experimental results and comparative analysis, evaluating the performance of MFCEN against existing methods. Finally, Section 5 concludes the paper with a summary of the findings and suggestions for future research.

2. Related works

2.1. CNNs

In recent years, the field of SISR has witnessed significant advancements in improving image details and enriching texture. Outstanding works appeared such as SRCNN [18], ESPCN [20], VDSR [21], LapSRN [34], EDSR [35], NLSA [24], and RCAN [36], among others. In comparison to traditional methods, SRCNN [18] pioneered the introduction of convolutional neural networks to address image super-resolution. It effectively extracted internal image features with just three convolutional layers and enhanced super-resolution performance through end-to-end training. EDSR [35] and DRCN [37], incorporating deeper and wider residual structures, better captured high-frequency details in images, progressively improving image super-resolution reconstruction performance through the use of residual blocks. LapSRN [34], MSRN [38], and others achieved increased resolution by constructing multi-scale features, enhancing the generality and adaptability of the networks. Despite significant progress in CNN-based image super-resolution, challenges persist, including insufficient generalization capability, higher computational complexity, and difficulties in capturing global relationships [39, 40].

2.2. Transformer

The transformer architecture has gained significant attention in the field of image super-resolution, particularly with the introduction of self-attention mechanisms. In comparison to traditional CNN and RNN methods, it effectively addresses long-range dependencies and supports parallel computation [41, 42], thereby enhancing network efficiency. In recent years, several transformer-based approaches have made noteworthy advancements in this domain. SwinIR [43], leveraging the Swin transformer architecture with a mobile window mechanism, successfully captures long-range dependencies and global context information. The texture transformer architecture treats the transformer as an attention module, improving image super-resolution quality through texture-aware loss functions.

Nevertheless, ViT structures still face challenges in image super-resolution tasks, such as high

computational costs and significant GPU memory usage. In response, Lu et al. [44] proposed the ESRT (efficient SR transformer) to explore the feasibility of using transformers in lightweight super-resolution tasks. Despite the commendable performance of existing networks, researchers actively seek efficient, real-time, and lightweight network architectures. This pursuit aims to optimize network structures, reduce parameters, and improve prediction speed without compromising performance.

2.3. Multi-scale feature

Despite the rapid development of vision transformers (ViTs) in the field of image super-resolution, their relatively shallow network architecture limits the effective reconstruction of local details. To address the shortcomings of the transformer structure, researchers introduce multi-scale feature extraction to comprehensively capture the spatial structure of images, enhancing the network's perception of both details and global context. In recent years, multi-scale features have been widely applied to various advanced visual tasks, demonstrating their effectiveness in improving model performance. For instance, CFNet [45] leverages multi-scale feature fusion to obtain the highest-level semantic features for dense prediction. HRNetV2 [46] cleverly integrates features at different scales, successfully capturing high-frequency details in images. SMSR [47] achieves adaptive feature detection and multi-scale feature fusion by effectively utilizing convolutional kernels of different sizes, thereby capturing high-frequency details in images and achieving high-quality image reconstruction. DLGSANet [48] introduces multi-head dynamic local self-attention and sparse global self-attention to dynamically extract local features and provide better self-attention for global feature exploration. However, most of these methods come at the cost of large parameter sizes and the high computational complexity associated with obtaining multi-scale features. Therefore, recent research has begun to explore approaches to obtaining rich feature representations at a smaller cost [49, 50]. The proposed lightweight multi-scale super-resolution method in this paper aims to maintain high performance while reducing network parameters and computational complexity to meet practical application constraints on computational resources.

3. Proposed method

In this manuscript, we introduce a novel lightweight, multi-scale feature extraction architecture designed for image super-resolution reconstruction. Multi-scale feature extraction in the deep feature model empowers models to capture intricate information across diverse levels of detail, enhancing their ability to reconstruct nuanced image structures and fortifying the resilience of the super-resolution process.

In Section 3.1, we outline the framework of the lightweight multi-scale super-resolution network, including shallow feature extraction, deep feature extraction, and image reconstruction. Then, in Section 3.2, we present the specifics related to the local feature capturer in the integrated multi-level feature module. Subsequently, in Section 3.3, we introduce the novel dynamic sparse attention mechanism proposed in the block-scale feature extractor. Finally, in Section 3.4, we provide a detailed exposition of the core components of the proposed global semantic capturer.

3.1. Network architecture

Our proposed multi-scale feature cooperative enhance network (MFCEN) is designed for the efficient upscaling of low-resolution images to their high-resolution counterparts.

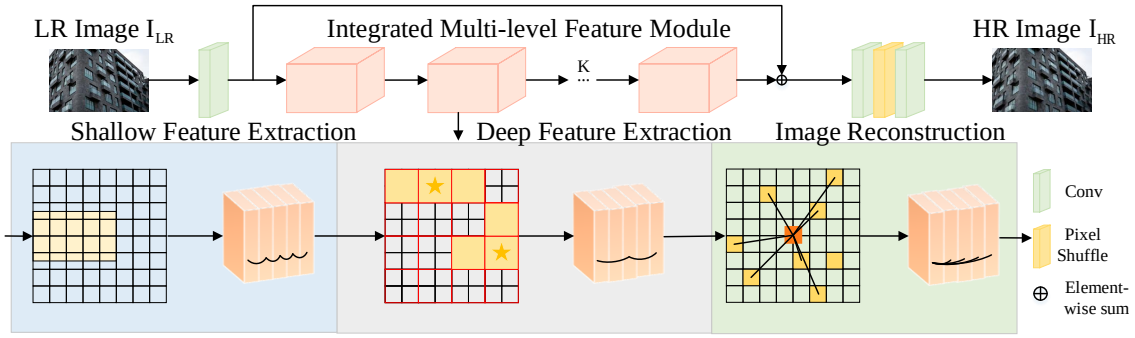

Figure 1. The framework of the lightweight multi-scale feature cooperative enhance network (MFCEN).

As depicted in Figure 1, MFCEN comprises three components: shallow feature extraction, deep feature extraction, and image reconstruction. The lower part of Figure 1 illustrates the three levels of deep feature extraction and their attention mechanisms: local, block-scale, and global feature extraction. The details of these mechanisms will be explained in the following sections.

To initiate the process, a 3×3 convolutional layer H_{SF} is applied to extract shallow features $F_{shallow} \in R^{H \times W \times C}$ from a given input low-resolution image (I_{LR}) :

$$F_{shallow} = H_{SF}\left(I_{LR}\right), \tag{3.1}$$

where H, W, and C are the feature height, width, and channels, respectively. This convolution layer not only efficiently extracts shallow features but also transforms the input from image space into a higher-dimensional feature space.

Subsequently, a deep feature extraction network is employed, consisting of K stacked integrated multi-level feature modules and residual connections, facilitating the extraction of profound features $F_{shallow}$. Each integrated multi-level feature module comprises a local feature capturer (LFC), block-scale feature extractor (BFE), and global semantic capturer (GSC), facilitating local propagation, multi-scale interactions, and global-scale engagement. This configuration constitutes a comprehensive aggregation building block. We will delve into the detailed description of each key element in Sections 3.2 to 3.4.

In essence, this process can be succinctly described as

$$F_{deep} = H_{DF}\left(F_{shallow}\right), \tag{3.2}$$

where the shallow and deep features are jointly learned to predict missing high-resolution images with rich details. The aggregated features are then reconstructed through PixelShuffle. Finally, we reconstruct the HR image I_{HR}

$$I_{HR} = H_{Rec}\left(F_{shallow} + F_{deep}\right), \tag{3.3}$$

where $H_{Rec}(\cdot)$ denotes the reconstruction module containing PixelShuffle which is used to upsample the fused feature.

3.2. Local feature capturer

In particular, the local feature capturer (LFC) is constructed by stacking depthwise separable convolutions, interleaved with a squeeze-and-excitation module (SE). This design facilitates the redistribution of channel attention while further extracting features and reducing computation complexity. The main goals of this module are to consolidate local contextual information and to reduce computation complexity. As depicted in Figure 2, the local feature capturer (LFC) is implemented as a stack of pointwise and depthwise convolutions with a squeeze-and-excitation module (SE) module between them to adaptively re-weight channel-wise features. This module aims to aggregate local contextual information as well as to increase the trainability of the network.

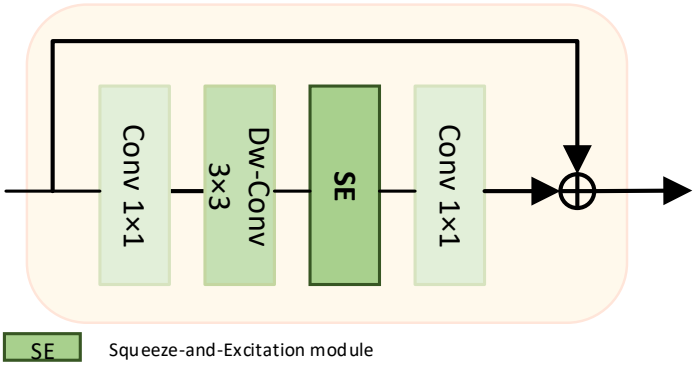

Figure 2. The framework of the local feature capturer.

3.3. Block-scale feature extractor

To address challenges associated with high memory consumption and computational costs, various strategies have been explored, such as restricting the attention operation to be inside local windows, axial stripes, or dilated windows. These techniques introduce sparse attention with different handcrafted modes to alleviate complexity. Sparse attention mechanisms in non-manual modes typically share a sampled subset of key-value pairs across all queries, which cannot interfere with each other. As shown in Figure 3, the block-scale feature extractor (BFE) in this paper employs an innovative dynamic sparse attention mechanism to achieve a more lightweight and flexible feature extraction at the block scale.

The key aspect of our approach involves filtering out the most irrelevant key-value pairs at the region level, retaining only a portion of routing regions. Then fine-grained token-to-token attention is jointly applied within these routing regions. We first construct a region-level association graph, followed by limiting the connections to the first k that are kept for each node. This effectively reduces computational complexity within each region. After determining the participating regions, we then apply token-to-token attention in the subsequent step. The design of this workflow aims to select regions relevant to the task, enhancing computational efficiency and reducing redundancy in the information processing flow.

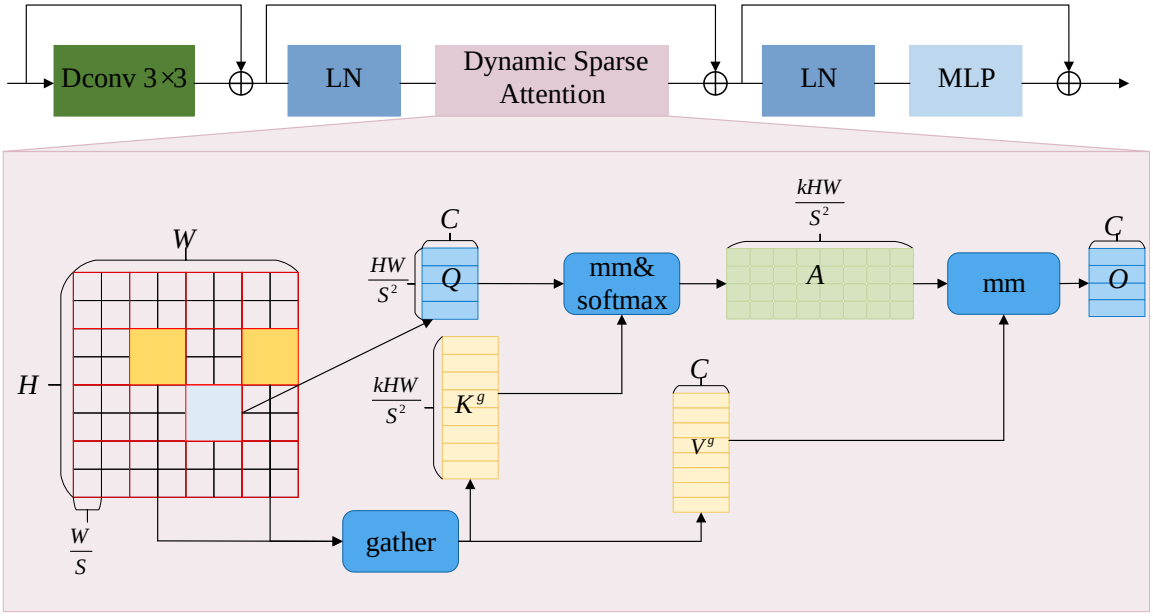

Figure 3. The framework of the block-scale feature extractor.

Specifically, we first partition the input image $X \in R^{H\times W\times C}$ into $S \times S$ different regions, each containing $\frac{HW}{S^2}$ feature vectors, transforming X into $X^r \in R^{S^2\times\frac{HW}{S^2}\times C}$. Then, through linear mapping, we obtain $Q, K, V \in R^{S^2\times\frac{HW}{S^2}\times C}$

$$Q = X^r W^q, K = X^r W^k, V = X^r W^v, \tag{3.4}$$

where W^q, W^k, and $W^v \in R^{C\times C}$ are the projection weights for query, key, and value.

Next, we construct a directed graph to determine the participation of each given region. First, we compute the average of Q and K within each region, obtaining $Q^r, K^r \in R^{S^2\times C}$. Then, we calculate the adjacency matrix of inter-regional correlations between Q^r and K^r :

$$A^r = Q^r (K^r)^T. \tag{3.5}$$

We then retain only the top K connections for each region to prune the correlation graph, saving the indices of the top K connections in the routing index matrix $I^r \in \mathbb{N}^{S^2\times k}$

$$I^r = topkIndex\,(A^r)\,, \tag{3.6}$$

where the i-th row of I^r contains the indices of the top K most relevant regions for the i-th region.

For each query token in region i, we focus on all key-value pairs in the union of the K routing regions with indices $I^r_{(i,1)}$, $I^r_{(i,2)}$,...,$I^r_{(i,k)}$. Specifically, we first gather the tensors for key and value

$$K^g = gather\,(K, I^r)\,, V^g = gather\,(V, I^r)\,, \tag{3.7}$$

where K^g and V^g are the gathered tensors for key and value. We then apply attention operation to the gathered K-V pairs and introduce a local context enhancement term $LEC\,(V)$,

$$O = Attention\,(Q, K^g, V^g) + LEC\,(V)\,. \tag{3.8}$$

Finally, by using patch embedding or patch merging, implicitly encoding relative position information with a 3×3 convolutional layer, and employing an MLP alongside an automatic query routing attention module, we create a block-scale feature module for feature extraction. This module is designed for modeling cross-position relationships and embedding position-wise information.

3.4. Global semantic capturer

Although the transformer model addresses the limitations of CNNs in capturing global features, its computational complexity significantly increases with the growth of spatial resolution. Hence, this paper introduces a powerful yet computationally efficient feature capturing model at a global scale. As shown in Figure 4, this module employs a cross-channel self-attention mechanism, coupled with a selectively forward-propagating network module, to capture global information while controlling computational complexity and reducing the required training parameters.

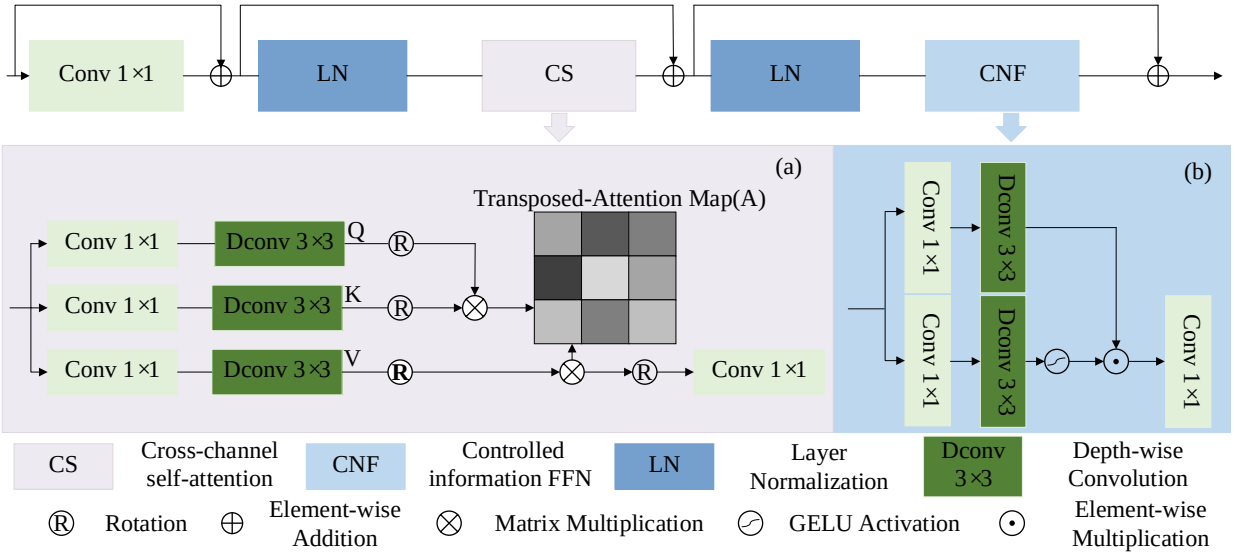

Figure 4. The framework of the global semantic capturer.

3.4.1. Cross-channel self-attention

We observe that the primary computational cost in the transformer arises from the self-attention operation, with time and memory complexity mainly attributed to the computation of key-query pairs. However, as spatial resolution increases, the computational complexity grows quadratically, making the use of spatial attention impractical in high-resolution images. Therefore, this module opts for cross-channel self-attention, implicitly encoding global contextual information by computing attention across channels.

Furthermore, before calculating the feature covariance to generate the global attention map, depth-wise convolution [51] is introduced to emphasize contextual information. Specifically, for an input image $X \in R^{H \times W \times C}$, it is first normalized to obtain $Y \in R^{H \times W \times C}$, followed by the generation of query (Q), key (K), and value (V). Pixel-level cross-channel context is aggregated using a 1×1 point-wise convolution, followed by encoding channel-level spatial context using a 3×3 depth-wise convolution:

$$Q = W_d^Q W_p^Q Y, K = W_d^K W_p^K Y, \\ V = W_d^V W_p^V Y, \tag{3.9}$$

where $W_p(\cdot)$ represents the 1×1 point-wise convolution, and $W_d(\cdot)$ corresponds to the 3×3 depth-wise convolution.

Subsequently, reshaping is applied to query and key, and a transposed-attention map $A \in \mathbb{R}^{C \times C}$ is generated through dot-product computation. In summary, this process can be described as

$$X = W_p Attention(Q, K, V) + X, \\ Attention(Q, K, V) = V \cdot Softmax(K \cdot Q/\alpha), \tag{3.10}$$

α is where a learnable scaling parameter to control the magnitude of the dot product of K and Q.

3.4.2. Controlled information FFN

For transformer features, the feed-forward network (FFN) performs the same operation for each pixel position. Two 1×1 convolutions are employed, one for expanding feature channels (by a factor of γ) and the other for reducing the channels back to the original input dimensions.

Additionally, a gating mechanism is introduced, choosing one path through element-wise multiplication and applying GELU non-linear activation. Following this, a depth-wise convolution is applied to encode information from spatially adjacent pixel positions, restoring the local structure of the image. Specifically

$$X = W_p^0 Gating(X) + X, \\ Gating(X) = \phi\left(W_d^1 W_p^1(LN(X))\right) \\ \odot W_d^2 W_p^2(LN(X)), \tag{3.11}$$

where $\odot$ denotes element-wise multiplication, ϕ represents the GELU non-linear activation, and LN is the normalization layer.

In summary, this module prioritizes cross-channel features, emphasizing spatial local context, while addressing channel-wise global information. It effectively controls information flow among layers, enabling each layer to focus on complementary fine details. The module, consisting of cross-channel self-attention and this mechanism, forms the core of the global attention module. It achieves efficient and precise global information capture while balancing computational requirements, thereby realizing lightweight image super-resolution reconstruction.

4. Experiment

In this section, we perform quantitative and qualitative evaluations to demonstrate the effectiveness of the proposed MFCEN on benchmarks.

4.1. Datasets

We have chosen DIV2K [52] as our training dataset, consisting of 800 high-resolution images. Serving as a benchmark in the field of image super-resolution, DIV2K provides a comparable

foundation for experiments, allowing direct comparisons with other state-of-the-art methods. Due to its inclusion of high-resolution images spanning various themes and scenes, DIV2K facilitates the learning of complex image structures by our model, offering comprehensive testing for the model's generalization performance in diverse scenarios.

For our test set, we considered Set5 [53], Set14 [54], B100 [55], Urban100 [56], and Manga109 [57], encompassing different domains and scenes, with three distinct upscaling factors (x2, x3, and x4). This selection ensures that our model undergoes thorough testing on diverse images while allowing our experimental results to be standardized for comparisons with other methodologies in the image super-resolution domain. To generate degraded data that simulates real-world image degradation, we employed a bicubic interpolation and a blur scale degradation model. This comprehensive dataset selection robustly supports our experiments, enabling our method to adapt effectively to various real-world scenarios.

4.2. Metrics

Our evaluation strategy encompasses a comprehensive set of robust metrics to assess the effectiveness of the proposed MFCEN model [35, 36, 58]. The chosen metrics offer a holistic understanding of the model's performance across various scales, ensuring an in-depth analysis.

We initiate with peak signal-to-noise ratio (PSNR), a classical metric for image fidelity. Applying PSNR at different scales allows us to evaluate the model's effectiveness in preserving image details. The structural similarity index (SSIM) serves as another pivotal metric, considering the structural information of images. Evaluating SSIM across different scales reveals the model's capability to maintain the integrity of image structures. Efficiency assessment generally involves scrutinizing the model parameters (Params). Comparative analysis across different scales ensures the model's adaptability to varying complexities and scenarios. Floating point operations per second (Flops) address the computational complexity of the model, especially crucial for lightweight models. Cross-scale evaluation of Flops enables an understanding of the model's execution efficiency under diverse resource constraints.

In summary, this comprehensive set of metrics provides a nuanced understanding of the proposed model's performance, ensuring its effectiveness in handling different scales and practical applications.

Furthermore, we optimize models by minimizing the L_1 loss through the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$). Our model is implemented based on Pytorch with 1 RTX 3090 (24 GB) GPU.

4.3. Comparison with state-of-the-art methods

4.3.1. Quantitative comparison

To validate the effectiveness of the methodology of this paper, we conducted a comparative analysis with several advanced lightweight algorithms and some classical algorithms. These algorithms were selected for their ability to maintain minimal parameter count and memory usage while striving to achieve high-quality super-resolution, aligning closely with the objectives of our model. The comparison includes SRCNN [18], VDSR [21], LapSRN [34], CRAN [59], SRMDNF [60], IMDN [61], PAN [62], RFDN [63], MAMNet [64], ShuffleMixer [65], RLFN [40], SPAN [67], and Hit-SIR [68].

The best results are bold and the second-best are underlined. As shown in Table 1, concerning the

PSNR metric, our proposed algorithm achieves better results at a scaling factor of x3 on most datasets, only slightly trailing behind the performance of another network on specific datasets. It is noteworthy that the reconstruction performance of our algorithm on the test set is still at a high level, despite the fact that it exhibits higher efficiency in terms of the number of parameters. This can be attributed to the straightforward and efficient structure adopted by our network, enabling comparable performance under relatively fewer parameters. Although our parameters increase slightly compared to SPAN at x2 magnification, we still achieve superior performance. Taking x4 super-resolution on the Urban100 dataset as an illustration, our algorithm demonstrates a parameter reduction of approximately 325 K compared to LapSRN, and around 227 K compared to IMDN. In terms of Flops, the number of Flops in MFCEN, the method proposed in this paper, is much smaller than networks such as ShuffleMixer. Subsequently, more specific experiments were conducted, which used scaling factors 2, 3, and 4 under equal conditions on five test sets, namely Set5, Set14, BSD100, Urban100, and Manga109, and the optimal results have been marked in Table 1. In terms of the SSIM metric, our algorithm excels on the test sets at various scaling factors, with particularly notable performance on the x3 scale test set. For instance, at a scaling factor of x3 on the Urban100 dataset, our algorithm outperforms MAMNet with a larger number of parameters and the latest HiT-SR.

Table 1. Quantitative comparisons with other SR methods.

Methods	Scale	Params (↓)	Flops (↓)	Set5		Set14		BSD100		Urban100		Manga109	
				PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (Datasets ↑)	SSIM (↑)
SRCNN	×2	57 K	52.7 G	36.66	0.9299	32.45	0.9607	31.36	0.8879	29.50	0.8946	35.60	0.9663
VDSR	×2	665 K	613 G	37.53	0.9587	33.13	0.9124	31.90	0.8960	30.76	0.9140	37.22	0.9729
LapSRN	×2	251 K	29.9 G	37.52	0.9591	32.99	0.9124	31.80	0.8952	30.41	0.9103	37.27	0.9740
CRAN	×2	1592 K	394 G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
SRMDNF	×2	1511 K	-	37.79	0.9601	33.32	0.9159	32.05	0.8985	31.33	0.9204	-	-
IMDN	×2	694 K	159 G	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
PAN	×2	261 K	70.5 G	38.00	0.9605	33.59	0.9181	32.18	0.8997	32.01	0.9273	38.70	0.9773
RFDN	×2	534 K	95.0 G	38.05	0.9606	33.72	0.9187	32.22	0.9000	32.33	0.9299	38.88	0.9773
MAMNet	×2	942 K	-	38.10	0.9601	33.90	0.9199	32.30	0.9007	32.94	0.9352	39.15	0.9772
ShuffleMixer	×2	394 K	91 G	38.01	0.9606	33.63	0.9180	32.17	0.8995	31.89	0.9257	-	-
RLFN	×2	527 K	115.4 G	38.07	0.9607	33.72	0.9187	32.22	0.9000	32.33	0.9299	-	-
SPAN	×2	431 K	-	38.08	0.9608	33.71	0.9183	32.22	0.9002	32.24	0.9294	38.94	0.9777
HiT-SIR	×2	772 K	209.9 G	**38.22**	0.9613	**33.91**	**0.9213**	**32.35**	0.9019	**33.02**	**0.9365**	**39.38**	**0.9782**
MFCEN (Ours)	×2	532 K	1.68 G	38.04	**0.9621**	33.48	0.9203	32.23	**0.9033**	32.18	0.9298	38.82	0.9771
SRCNN	×3	57 K	52.7 G	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989	30.48	0.9117
VDSR	×3	665 K	613 G	33.66	0.9213	29.77	0.8314	28.82	0.7976	27.14	0.8279	32.01	0.9310
LapSRN	×3	-	-	33.82	0.9226	29.87	0.8320	28.82	0.7974	27.07	0.8281	32.21	0.9352
CRAN	×3	1592 K	119 G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
SRMDNF	×3	1528 K	-	34.12	0.9254	30.04	0.8382	28.97	0.8025	27.57	0.8398	33.57	0.9442
IMDN	×3	703 K	72 G	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
PAN	×3	261 K	39.0 G	34.40	0.9271	30.36	0.8423	29.11	0.8050	28.11	0.8511	33.61	0.9448
RFDN	×3	541 K	42.2 G	34.41	0.9273	30.34	0.8420	29.09	0.8050	28.21	0.8525	33.67	0.9449
MAMNet	×3	1127 K	-	34.61	0.9281	30.54	0.8459	29.25	0.8082	28.82	0.8648	34.14	0.9472
ShuffleMixer	×3	415 K	43 G	34.40	0.9272	30.37	0.8423	29.12	0.8051	28.08	0.8498	33.69	0.9448
RLFN	×3	-	-	34.42	0.9278	30.33	0.8419	29.10	0.8051	28.21	0.8525	-	-
HiT-SR	×3	780 K	94.2 G	**34.72**	0.9298	**30.62**	**0.8474**	**29.27**	0.8101	28.93	0.8673	**34.40**	**0.9496**
MFCEN (Ours)	×3	596 K	1.86 G	34.39	**0.9298**	30.08	0.8439	**29.14**	**0.8112**	**32.98**	**0.9368**	33.57	0.9438
SRCNN	×4	57 K	52.7 G	30.48	0.8628	27.49	0.7503	26.90	0.7101	24.52	0.7221	27.66	0.8505
VDSR	×4	665 K	613 G	31.35	0.8838	28.01	0.7674	27.29	0.7251	25.18	0.7524	28.83	0.8809
LapSRN	×4	813 K	149.4 G	31.54	0.8852	28.09	0.7700	27.32	0.7275	25.21	0.7562	29.09	0.8900
CRAN	×4	1592 K	91 G	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
SRMDNF	×4	1552 K	-	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7331	30.09	0.9024
IMDN	×4	715 K	41 G	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
PAN	×4	272 K	28.2 G	32.13	0.8948	28.61	0.7822	27.59	0.7363	26.11	0.7854	30.51	0.9095
RFDN	×4	550 K	23.9 G	32.24	0.8952	28.61	0.7819	27.57	0.7360	26.11	0.7858	30.58	0.9089
MAMNet	×4	1090 K	-	32.42	0.8972	28.77	0.7854	27.70	0.7406	26.59	0.8013	30.94	0.9142
ShuffleMixer	×4	411 K	28 G	32.21	0.8953	28.66	0.7827	27.61	0.7366	26.08	0.7835	30.65	0.9093
RLFN	×4	543 K	29.8 G	32.24	0.8952	28.62	0.7813	27.60	0.7364	26.17	0.7877	-	-
SPAN	×4	498 K	-	32.20	0.8953	28.66	0.7834	27.62	0.7374	26.18	0.7879	30.66	0.9103
HiT-SR	×4	772 K	53.8 G	**32.51**	**0.8991**	**28.84**	0.7873	**27.73**	**0.7424**	**26.71**	**0.8045**	**31.23**	**0.9176**
MFCEN (Ours)	×4	488 K	1.92 G	32.12	0.8974	28.31	**0.7875**	27.61	0.744	26.07	0.7867	30.42	0.9081

We can intuitively see that MFCEN has good results on all scaling factors. The main reason for this is that while using the integrated multi-level feature module to provide the network with a rich representation of feature information, the extraction of key feature information in the features outperforms the existing mainstream models in terms of performance. In summary, MFCEN not only exhibits superior performance at different scales but also features fewer network parameters and low Flops.

4.3.2. Visual comparison

For the analysis of the visual effect of the final reconstruction of the model, figures in the test set Urban100 were selected for the experiments, and a comparison of the effects presented after MFCEN super-resolution reconstruction is shown in Figure 5.

From the final reconstruction effect, it can be seen that many models are unable to reconstruct the details, which may ignore the changes of local details, and it may appear that the image is too smooth and has lost the original details and texture information. When the image magnification is high, some algorithms show jagged lines in the edge part of the image. In contrast, MFCEN can accurately reconstruct a relatively clear super-resolution image, taking into account the details of the image texture while focusing on the overall effect, even if the scale factor increases. The lightweight model MFCEN proposed in this paper has an advantage in the visual effect.

4.4. Ablation studies

In this section, ablation experiments are carried out to validate the effectiveness of the different modules proposed in the algorithm of this paper and to determine whether it is necessary to use all three modules simultaneously to show the best performance. In this case, the network without the addition of the LFC, BFE, and GSC modules is taken as the baseline network Baseline. F(a) is the baseline model, and F(b)–F(d) are the network models obtained after the addition of the BFE with GSC, LFC with GSC, and LFC with BFE modules to the baseline structure, respectively.

As can be seen from the results in Table 2, the reconstruction performance is significantly improved when the LFC module is used instead of the regular convolution for local feature extraction and combined with the BFE module for block-scale feature extraction. In addition, the introduction of the squeeze-and-excitation (SE) module [66] with the dynamic sparse attention mechanism effectively reduces the number of parameters in the network. Compared to the baseline structure, the average PSNR metric of our algorithm is improved by 0.0758 dB, indicating that our method is more effective in local and block-scale information extraction. The application of LFC and BFE modules also leads to an improvement in the SSIM metric of the algorithm by 0.0029. In addition, the number of parameters is decreased from 756 K to 530 K, which achieves the lightweight quality of the algorithm while maintaining its reconstruction performance.

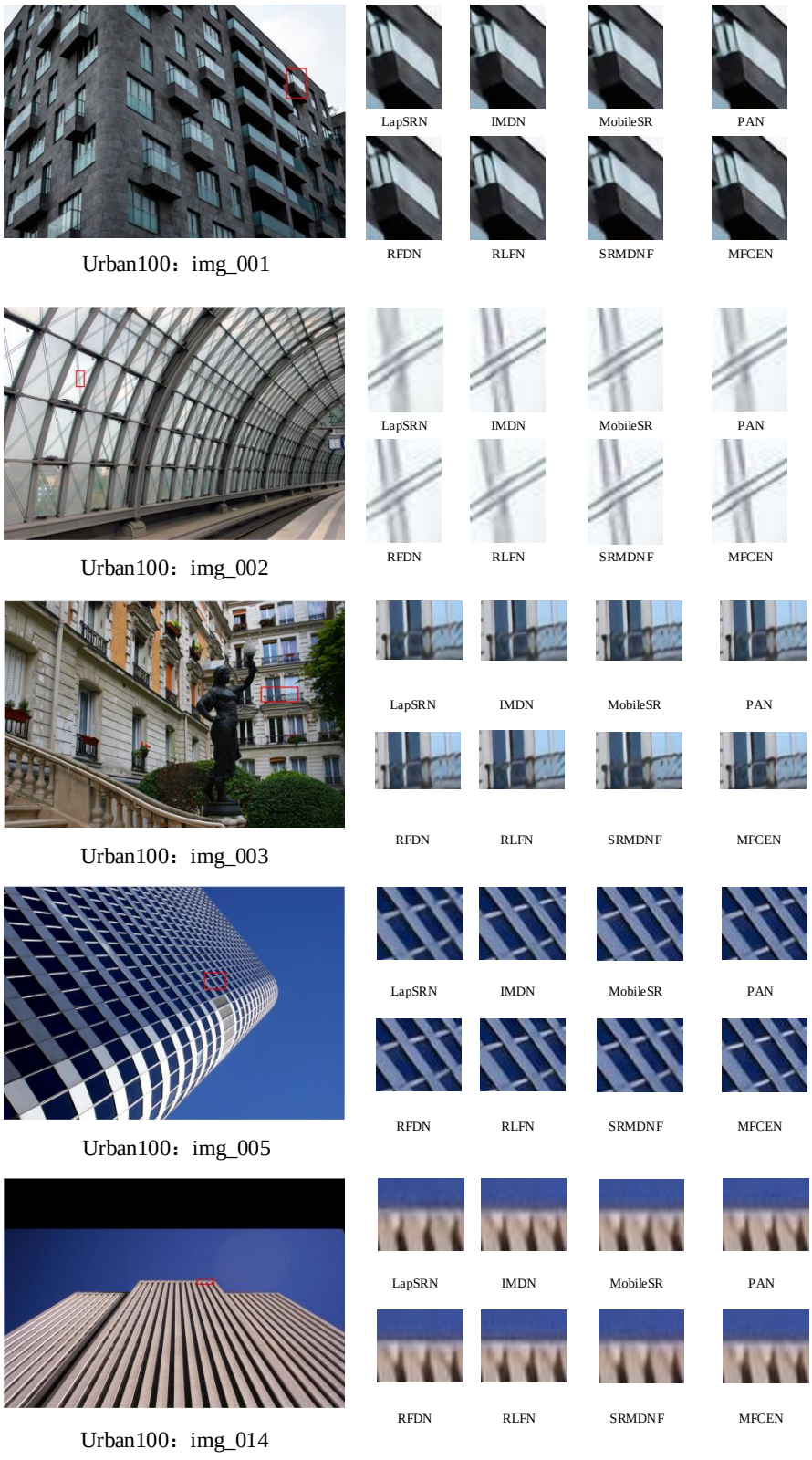

Figure 5. Results of the experiment on Urban100.

Table 2. Ablation study of MFCEN.

	LFC	BFE	GSC	Params	Flops	PSNR	SSIM
F(a)				756 K	5.68 G	26.0194	0.7844
F(b)		✓	✓	688 K	5.36 G	26.0952	0.7873
F(c)	✓		✓	692 K	4.36 G	26.1346	0.7901
F(d)	✓	✓		530 K	3.38 G	26.2346	0.7912

When the BFE and GSC modules are introduced, the PSNR metric of the algorithm is significantly improved by 0.1969 dB concerning the baseline model. The cross-channel self-attention mechanism in the GSC module can accurately extract the global channel information and merge it with the block-scale features extracted by the BFE module to adaptively adjust the feature maps to reconstruct the relevant features while blocking the irrelevant reconstructed features. Moreover, the number of parameters of the algorithm is reduced by the controlled information forward feedback network in GSC with dynamic sparse self-attention in BFE.

Compared to the baseline network, the PSNR of the algorithm is significantly improved by 0.2152 dB with the introduction of the LFC and GSC modules, and the number of parameters of the model is significantly reduced by using the depthwise separable convolution. In addition, the excellent local feature capture ability based on CNN and the powerful global semantic information capture ability based on ViTs structure make the reconstruction results significantly improved. The data shows that the SSIM of the model is also improved. Therefore, the LFC and GSC modules can improve the reconstruction performance without increasing the number of parameters.

The use of the LFC and BFE modules significantly reduces the number of parameters. The GSC module, through its cross-channel self-attention mechanism and controlled information forward feedback network, accurately extracts and integrates global semantic information. When the GSC module is combined with the LFC and BFE modules, the algorithm shows a significant improvement in PSNR and SSIM, ensuring that the model remains lightweight while still delivering excellent reconstruction performance.

5. Conclusions

We have proposed a lightweight multi-scale feature cooperative enhancement network (MFCEN) specifically designed for single-image super-resolution. The network is composed of integrated multi-scale feature modules (IMFMs), each consisting of a local feature capturer, block-Scale feature extractor, and global semantic capturer. In the domain of local feature extraction, we enhance the efficiency of capturing local features by fusing deep convolution with a channel attention mechanism. The introduction of dynamic sparse attention effectively filters feature information at the block scale, improving computational efficiency. Additionally, incorporating cross-channel attention and a controlled information feed-forward network enhances the extraction of global semantic information. Compared to existing image super-resolution reconstruction methods, our network conducts multi-scale extraction in three dimensions, which captures useful information more comprehensively for subsequent high-resolution (HR) image reconstruction. Experimental results demonstrate that MFCEN successfully reconstructs higher-quality HR images and significantly reduces parameter count and computational complexity, showcasing the feasibility of SR networks in real-life

applications. This research achieves a significant breakthrough in single-image super-resolution, providing an efficient and viable solution for practical applications.

Despite these advancements, the multi-scale feature cooperative enhancement network (MFCEN) still exhibits some limitations. For instance, its performance in cross-domain applications has not yet met expectations. Experimental results indicate that, particularly when applied to remote sensing and underwater images, the model may require targeted adjustments and optimizations to enhance its performance. To improve MFCEN's effectiveness in these specific areas, future research should focus on further refining the model, ensuring better adaptability and robustness across various real-world scenarios.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Author Contributions

Conceptualization, Jiange Liu and Yu Chen; Data curation, Xin Dai and Li Cao; Formal analysis, Yu Chen; Funding acquisition, Jiange Liu and Qingwu Li; Investigation, Yu Chen, Li Cao and Xin Dai; Methodology, Li Cao and Yu Chen; Project administration, Jiange Liu; Resources, Jiange Liu, Yu Chen, and Qingwu Li; Software, Xin Dai and Qingwu Li; Supervision, Jiange Liu and Qingwu Li; Validation, Yu Chen, Li Cao and Xin Dai; Visualization, Li Cao and Yu Chen; Writing—original draft, Yu Chen and Xin Dai; Writing—review and editing, Li Cao and Yu Chen. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors are thankful for the funding from the Science and Technology Project of State Grid Jiangsu Electric Power Co., Ltd. (J2023061).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. H. Guan, Y. Hu, J. Zeng, C. Zuo, Q. Chen, Super-resolution imaging by synthetic aperture with incoherent illumination, *Comput. Imaging VII*, **12523** (2023), 100–104.
2. H. M. Patel, V. M. Chudasama, K. Prajapati, K. P. Upla, K. Raja, R. Ramachandra, et al., ThermISRnet: An efficient thermal image super-resolution network, *Opt. Eng.*, **60** (2021), 073101. https://doi.org/10.1117/1.OE.60.7.073101
3. D. Qiu, Y. Cheng, X. Wang, Medical image super-resolution reconstruction algorithms based on deep learning: A survey, *Comput. Methods Programs Biomed.*, **238** (2023), 107590. https://doi.org/10.1016/j.cmpb.2023.107590

4. H. Yang, Z. Wang, X. Liu, C. Li, J. Xin, Z. Wang, Deep learning in medical image super resolution: A review, *Appl. Intell.*, **53** (2023), 20891–20916. https://doi.org/10.1007/s10489-023-04566-9
5. C. Wang, J. Jiang, K. Jiang, X. Liu, SPADNet: Structure prior-aware dynamic network for face super-resolution, *IEEE Trans. Biom. Behav. Identity Sci.*, **6** (2024), 326–340. https://doi.org/10.1109/TBIOM.2024.3382870
6. C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, M. Norouzi, Image super-resolution via iterative refinement, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2022), 4713–4726. https://doi.org/10.1109/TPAMI.2022.3204461
7. G. Bhat, M. Danelljan, L. Van Gool, R. Timofte, Deep burst super-resolution, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2021), 9205–9214. https://doi.org/10.1109/CVPR46437.2021.00909
8. A. Lugmayr, M. Danelljan, L. Van Gool, R. Timofte, Srflow: Learning the super-resolution space with normalizing flow, in *Computer Vision – ECCV 2020*, Springer, (2020), 715–732. https://doi.org/10.1007/978-3-030-58558-7_42
9. K. Zhang, L. Van Gool, R. Timofte, Deep unfolding network for image super-resolution, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2020), 3214–3223. https://doi.org/10.1109/CVPR42600.2020.00328
10. X. Kong, H. Zhao, Y. Qiao, C. Dong, Classsr: A general framework to accelerate super-resolution networks by data characteristic, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2021), 12011–12020. https://doi.org/10.1109/CVPR46437.2021.01184
11. X. Wang, L. Xie, C. Dong, Y. Shan, Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data, in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, IEEE, (2021), 1905–1914. https://doi.org/10.1109/ICCVW54120.2021.00217
12. Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, et al., Closed-loop matters: Dual regression networks for single image super-resolution, in *020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2020), 5406–5415. https://doi.org/10.1109/CVPR42600.2020.00545
13. Z. Yue, J. Wang, C. C. Loy, Resshift: Efficient diffusion model for image super-resolution by residual shifting, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., **36** (2024), 13294–13307.
14. L. Sun, J. Dong, J. Tang, J. Pan, Spatially-adaptive feature modulation for efficient image super-resolution, in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2023), 13144–13153. https://doi.org/10.1109/ICCV51070.2023.01213
15. Z. Du, D. Liu, J. Liu, J. Tang, G. Wu, L. Fu, Fast and memory-efficient network towards efficient image super-resolution, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, (2022), 852–861. https://doi.org/10.1109/CVPRW56347.2022.00101
16. J. L. Harris, Diffraction and resolving power, *J. Opt. Soc. Am.*, **54** (1964), 931–936. https://doi.org/10.1364/JOSA.54.000931

17. D. C. Lepcha, B. Goyal, A. Dogra, V. Goyal, Image super-resolution: A comprehensive review, recent trends, challenges and applications, *Inf. Fusion*, **91** (2023), 230–260. https://doi.org/10.1016/j.inffus.2022.10.007

18. C. Dong, C. C. Loy, K He, X. Tang, Image super-resolution using deep convolutional networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **38** (2015), 295–307. https://doi.org/10.1109/TPAMI.2015.2439281

19. C. Dong, C. C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in *Computer Vision – ECCV 2016*, Springer, (2016), 391–407. https://doi.org/10.1007/978-3-319-46475-6_25

20. W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, et al., Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2016), 1874–1883. https://doi.org/10.1109/CVPR.2016.207

21. J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2016), 1646–1654. https://doi.org/10.1109/CVPR.2016.182

22. L. Zhang, X. Li, D. He, F. Li, Y. Wang, Z. Zhang, RRSR: Reciprocal reference-based image super-resolution with progressive feature alignment and selection, in *Computer Vision – ECCV 2022*, Springer, (2022), 648–664. https://doi.org/10.1007/978-3-031-19800-7_38

23. Y. Yang, W. Ran, H. Lu, Rddan: A residual dense dilated aggregated network for single image deraining, in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, (2020), 1–6. https://doi.org/10.1109/ICME46284.2020.9102945

24. Y. Mei, Y. Fan, Y. Zhou, Image Super-Resolution with Non-Local Sparse Attention, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2021), 3516–3525. https://doi.org/10.1109/CVPR46437.2021.00352

25. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2021), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

26. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, preprint, arXiv:2010.04159.

27. X. Zhu, H. Hu, S. Lin, J. Dai, Deformable ConvNets V2: More deformable, better results, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2019), 9300–9308. https://doi.org/10.1109/CVPR.2019.00953

28. S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2021), 6877–6886. https://doi.org/10.1109/CVPR46437.2021.00681

29. M. Zheng, P. Gao, R. Zhang, K. Li, X. Wang, H. Li, et al., End-to-end object detection with adaptive clustering transformer, preprint, arXiv:2011.09315.

30. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, (2021), 10347–10357.

31. P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, et al., Multi-scale vision longformer: A new vision transformer for high-resolution image encoding, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2021), 2978–2988. https://doi.org/10.1109/ICCV48922.2021.00299

32. J. Fang, H. Lin, X. Chen, K. Zeng, A hybrid network of cnn and transformer for lightweight image super-resolution, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, (2022), 1102–1111. https://doi.org/10.1109/CVPRW56347.2022.00119

33. G. Gao, Z. Wang, J. Li, W. Li, Y. Yu, T. Zeng, Lightweight bimodal network for single-image super-resolution via symmetric CNN and recursive transformer, preprint, arXiv:2204.13286.

34. W. S. Lai, J. B. Huang, N. Ahuja, M. H. Yang, Deep laplacian pyramid networks for fast and accurate super-resolution, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2017), 5835–5843. https://doi.org/10.1109/CVPR.2017.618

35. B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, Enhanced deep residual networks for single image super-resolution, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, (2017), 1132–1140. https://doi.org/10.1109/CVPRW.2017.151

36. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in *Computer Vision – ECCV 2018*, Springer, (2018), 294–310. https://doi.org/10.1007/978-3-030-01234-2_18

37. J. Kim, J. K. Lee, K. M. Lee, Deeply-recursive convolutional network for image super-resolution, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2016), 1637–1645. https://doi.org/10.1109/CVPR.2016.181

38. J. Li, F. Fang, K. Mei, G. Zhang, Multi-scale residual network for image super-resolution, in *Computer Vision – ECCV 2018*, Springer, (2018), 527–542. https://doi.org/10.1007/978-3-030-01237-3_32

39. F. Zhu, Q. Zhao, Efficient single image super-resolution via hybrid residual feature learning with compact back-projection network, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, (2019), 2453–2460. https://doi.org/10.1109/ICCVW.2019.00300

40. F. Kong, M. Li, S. Liu, D. Liu, J. He, Y. Bai, et al., Residual local feature network for efficient super-resolution, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, (2022), 765–775. https://doi.org/10.1109/CVPRW56347.2022.00092

41. J. Yang, S. Shen, H. Yue, K. Li, Implicit transformer network for screen content image continuous super-resolution, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., **34** (2021), 13304–13315.

42. J. Li, S. Zhu, Channel-spatial transformer for efficient image super-resolution, in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, (2024), 2685–2689. https://doi.org/10.1109/ICASSP48485.2024.10446047
43. J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, SwinIR: Image restoration using swin transformer in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, IEEE, (2021), 1833–1844. https://doi.org/10.1109/ICCVW54120.2021.00210
44. Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, T. Zeng, Transformer for single image super-resolution, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, (2022), 456–465. https://doi.org/10.1109/CVPRW56347.2022.00061
45. X. Deng, Y. Zhang, M. Xu, S. Gu, Y. Duan, Deep coupled feedback network for joint exposure fusion and image super-resolution, *IEEE Trans. Image Process.*, **30** (2021), 3098–3112. https://doi.org/10.1109/TIP.2021.3058764
46. J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, et al., Deep high-resolution representation learning for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2021), 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686
47. L. Wang, X. Dong, Y. Wang, X. Ying, Z. Lin, W. An, et al., Exploring sparsity in image super-resolution for efficient inference, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2021), 4915–4924. https://doi.org/10.1109/CVPR46437.2021.00488
48. X. Li, J. Dong, J. Tang, J. Pan, DLGSANet: Lightweight dynamic local and global self-attention networks for image super-resolution, in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2023), 12746–12755. https://doi.org/10.1109/ICCV51070.2023.01175
49. W. Deng, H. Yuan, L. Deng, Z. Lu, Reparameterized residual feature network for lightweight image super-resolution, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, (2023), 1712–1721. https://doi.org/10.1109/CVPRW59228.2023.00172
50. X. Zhang, H. Zeng, S. Guo, L. Zhang, Efficient long-range attention network for image super-resolution, in *Computer Vision – ECCV 2022*, Springer, (2022), 649–667. https://doi.org/10.1007/978-3-031-19790-1_39
51. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications, preprint, arXiv:1704.04861.
52. R. Timofte, S. Gu, J. Wu, L. Van Gool, L. Zhang, M. H. Yang, et al., Ntire 2018 challenge on single image super-resolution: Methods and results, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2018), 965–96511. https://doi.org/10.1109/CVPRW.2018.00130
53. M. Bevilacqua, A. Roumy, C. Guillemot, M. L. Alberi-Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding, in *Proceedings of the 23rd British Machine Vision Conference (BMVC)*, BMVA Press, (2012), 1–10.
54. R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in *Curves and Surfaces. Curves and Surfaces 2010*, Springer, (2012), 711–730. https://doi.org/10.1007/978-3-642-27413-8_47

55. D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in *Proceedings Eighth IEEE International Conference on Computer Vision (ICCV)*, IEEE, (2001), 416–423. http://doi.org/10.1109/ICCV.2001.937655

56. J. B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, (2015), 5197–5206. http://doi.org/10.1109/CVPR.2015.7299156

57. Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, et al., Sketch-based manga retrieval using manga109 dataset, *Multimedia Tools Appl.*, **76** (2017), 21811–21838. https://doi.org/10.1007/s11042-016-4020-z

58. Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2018), 2472–2481.

59. Y. Zhang, D. Wei, C. Qin, H. Wang, H. Pfister, Y. Fu, Context reasoning attention network for image super-resolution, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, (2021), 4278–4287. https://doi.org/10.1109/ICCV48922.2021.00424

60. K. Zhang, W. Zuo, L. Zhang, Learning a single convolutional super-resolution network for multiple degradations, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, (2018), 3262–3271. https://doi.org/10.1109/CVPR.2018.00344

61. Z. Hui, X. Gao, Y. Yang, X. Wang, Lightweight image super-resolution with information multi-distillation network, in *Proceedings of the 27th ACM International Conference on Multimedia*, Association for Computing Machinery, (2019), 2024–2032. https://doi.org/10.1145/3343031.3351084

62. H. Zhao, X. Kong, J. He, Y. Qiao, C. Dong, Efficient image super-resolution using pixel attention, in *Computer Vision – ECCV 2020 Workshops*, Springer, (2020), 56–72. https://doi.org/10.1007/978-3-030-67070-2_3

63. J. Liu, J. Tang, G. Wu, Residual feature distillation network for lightweight image super-resolution, in *Computer Vision – ECCV 2020 Workshops*, Springer, (2020), 41–55. https://doi.org/10.1007/978-3-030-67070-2_2

64. J. H. Kim, J. H. Choi, M. Cheon, J. S. Lee, MAMNet: Multi-path adaptive modulation network for image super-resolution, *Neurocomputing*, **402** (2020), 38–49. https://doi.org/10.1016/j.neucom.2020.03.069

65. L. Sun, J. Pan, J. Tang, Shufflemixer: An efficient convnet for image super-resolution, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., **35** (2022), 17314–17326.

66. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2018), 7132–7141.

67. C. Wan, H. Yu, Z. Li, Y. Chen, Y. Zou, Y. Liu, et al., Swift parameter-free attention network for efficient super-resolution, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2024), 6246–6256.

68. X. Zhang, Y. Zhang, F. Yu, HiT-SR: Hierarchical transformer for efficient image super-resolution, preprint, arXiv:2407.05878.

AIMS Press